Journal of Forestry Research (2008) 19(4):293−297
DOI 10.1007/s11676-008-0052-1

RESEARCH PAPER

Analysis of synonymous codon usage in chloroplast genome of *Populus alba*

ZHOU Meng [1], LONG Wei [2], LI Xia [1*]

[1] *College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150086, P. R. China*
[2] *The Key Laboratory of Forestry Genetics and Engineering of State Forestry Administration & Jiangsu Provincial, Nanjing Forestry University, Nanjing 210037, P. R. China*

Abstract: The pattern of codon usage in the chloroplast genome of *Populus alba* was investigated. Correspondence analysis (a commonly used multivariate statistical approach) and method of effective number of codons (ENc)-plot were conducted to analyze synonymous codon usage. The results of correspondence analysis showed that the distribution of genes on the major axis was significantly correlated with the frequency of use of G+C in synonymously variable third position of sense codon (GC_{3S}), (r=0.349), and the positions of genes on the axis 2 and axis 3 were significantly correlated with CAI (r=-0.348, p<0.01 and r=0.602, p<0.01). The ENc for most genes was similar to that for the expected ENc based on the GC_{3S}, but several genes with low EN_C values were lying below the expected curve. All of these data indicated that codon usage was dominated by a mutational bias in chloroplast genome of *P. alba*. The selection in nature for translational efficiency only played a minor role in shaping codon usage in the chloroplast genome of *P. alba*.

Keywords: codon usage; chloroplast; GC content; *Populus alba*

Introduction

Sixty-four codons are found in the universal genetic code, which encode 20 different amino acids in the organism world. Owing to the degeneracy of the genetic code, each amino acid may be coded by two or more codons (synonymous codons). Non-random codon usage, or codon bias is a common phenomenon in a wide variety of organisms, including prokaryotes, animals and plants (Akashi 2001; Duret 2002; Bonitz, et al. 1980). Synonymous codon usage varies widely between genomes and also between genes within genomes (Wang 2007). The variations in interspecific codon usage and intragenomic codon usage are primarily due to directional mutation pressure on DNA sequences and natural selection affecting gene translation (Lu et al. 2005). Several authors previously suggested that the patterns of codon usage had shown different features between monocot and dicot species in nuclear genomes (Kawabe et al. 2003; Wang et al. 2007). The numerous analysis reports on synonymous codon usage bias have been mainly focused on nuclear genomes. However, only few mentions have been made of this analysis on organelles (Morton 1998, 1999, 2003; Morton et al. 2000). In the plant chloroplast genomes, selection in nature is found to be weaker in *Pinus thunbergii*, but there is evidence that an intermediate level of selection exists in the liverwort (*Marchantia polymorpha*), (Morton 1998). In the present study, we examined the pattern of synonymous codon usage in the chloroplast genome of *Populus alba*. The main purpose of this study is to investigate the codon usage pattern in this organelle.

Materials and methods

Sequence data

The complete chloroplast genome sequence from *P. alba* (NC_008235) was obtained from GenBank. Using the information in the Genbank file, all protein coding, ORF, and ycf (Hallick et al. 1994) or conserved open reading frames (Hallick et al. 1994) sequences greater than 300 nucleotides in length, were extracted directly to avoid sampling bias in codon usage calculations (Wright 1990). A total of 57 genes were combined for codon usage analysis.

Indices of codon usage

The effective number of codons (ENc) used in a gene is a simple

Foundation project: This study was supported by the National High Tech Development Project of China, the 863 Program (Grant Nos. 2007AA02Z329) and the National Natural Science Foundation of China (Grant Nos. 20060213024).

Received: 2008-06-10; Accepted: 2008-07-17

The online version is available at http://www.springerlink.com

Biography: ZHOU Meng (1981-), male, Master, Assistant in Harbin Medical University, Harbin 150086, P. R. China. (E-mail: biofomeng@hotmail.com)

* Corresponding author: LI Xia (E-mail: lixia@ems.hrbmu.edu.cn)

Responsible editor: Zhu Hong

measure of codon bias (Wright 1990). This is a measure of no uniformity of usage within synonymous groups of codons. EN_C values can vary from 20 (extreme bias where only one codon is used per amino acid) to 61 (random codon usage). Relative Synonymous Codon Usage (RSCU) is the observed frequency of a codon divided by the frequency expected if there is uniform usage within synonymous codon groups (Sharp et al. 1986). If all synonymous codons coding the same mino acid were used equally, RSCU values were close to 1.0, indicating a lack of bias. The index of GC_{3S} is the frequency of use of G+C in synonymously variable third positions of sense codon (i.e., excluding Met, Trp and termination codons). The "Codon Adaptation Index" (CAI) uses a reference set of highly expressed genes from a species to assess the relative merits of each codon, and a score for a gene is calculated from the frequency of use of all codons in that gene. The index assesses the extent to which selection has been effective in molding the pattern of codon usage. In that respect, CAI is useful for predicting the level of expression of a gene (Sharp et al. 1987). The CAI value for every gene was calculated relative to the *psbA* gene of the same genome (Moton 1998). These indices of codon usage bias were calculated for each gene in the data set using the program CodonW (version 1.4.2, http://codonw.sourceforge.net/).

Correspondence analysis

Correspondence analysis (COA) has become the method of choice for multivariate statistical analysis of codon usage patterns (Grantham et al. 1980; Shields et al. 1987; Sharp et al. 1989). Since there are a total of 59 synonymous codons (61 sense codons, less the unique methionine and tryptophan codons), this analysis partitions the variation along 59 orthogonal axes, with 41 degrees of freedom. The first axis is the one that captures most of the variation in codon usage, with each subsequent axis explaining a diminishing amount of the variance. In contrast to other types of variance component analysis, such as Principal Component Analysis (PCA), correspondence analysis has the advantage of not only to show the distribution of genes in the multidimensional space, but also to show the corresponding distribution of synonymous codons. Correspondence analysis was primarily designed for use with data tables containing counts, e.g., numbers of synonymous codons, whereas PCA is a general method of data reduction that is more suitable for continuous measurement data (Perriere et al. 2002). Correspondence analysis of RSCU was also performed using CodonW.

Statistical analysis

All correlations used are based on the nonparametric Spearman's rank correlation analysis method wrapped in the multi-analysis software SPSS Version 12.0. By using this measure of association, it is not essential to make any distributional assumptions of the underlying data.

Results

The codon usage for 57 chloroplast genes from *Populus alba* is presented in Table 1. There is a general excess of A- and U-ending codons. For every amino acid, an A- and U-ending codon is available. Terminal codon prefers to use UAA more than others often. For those amino acids, there is substantial (and statistically significant) no uniformity in synonymous codon usage (most easily seen by examination of the RSCU values). This reflects a mutational bias towards A+T, which in the absence of other selection pressures would be expected to increase the RSCU values for synonymous A+U ending codons to greater than 1.

Table 1. Summary of codon usage in the chloroplast genome of *Populus alba*

Amino acid	Codon	N	RSCU	Amino acid	Codon	N	RSCU	Amino acid	Codon	N	RSCU	Amino acid	Codon	N	RSCU
Phe	UUU	892	1.31	Ser	UCU	520	1.67	Tyr	UAU	725	1.65	Cys	UGU	190	1.39
	UUC	468	0.69		UCC	313	1.01		UAC	154	0.35		UGC	83	0.61
Leu	UUA	793	1.87		UCA	373	1.20	TER	UAA	31	1.63	TER	UGA	13	0.68
	UUG	518	1.22		UCG	172	0.55		UAG	13	0.68	Trp	UGG	428	1.00
	CUU	542	1.28	Pro	CCU	385	1.57	His	CAU	444	1.50	Arg	CGU	306	1.27
	CUC	156	0.37		CCC	187	0.76		CAC	148	0.50		CGC	106	0.44
	CUA	366	0.86		CCA	277	1.13	Gln	CAA	668	1.54		CGA	326	1.36
	CUG	167	0.39		CCG	131	0.53		CAG	199	0.46		CGG	91	0.38
Ile	AUU	1026	1.48	Thr	ACU	459	1.55	Asn	AAU	925	1.54	Ser	AGU	374	1.20
	AUC	397	0.57		ACC	224	0.76		AAC	276	0.46		AGC	112	0.36
	AUA	650	0.94		ACA	381	1.29	Lys	AAA	986	1.49	Arg	AGA	451	1.88
Met	AUG	544	1.00		ACG	118	0.40		AAG	340	0.51		AGG	161	0.67
Val	GUU	447	1.42	Ala	GCU	567	1.82	Asp	GAU	792	1.58	Gly	GGU	516	1.26
	GUC	155	0.49		GCC	197	0.63		GAC	210	0.42		GGC	178	0.43
	GUA	472	1.50		GCA	348	1.12	Glu	GAA	963	1.49		GGA	657	1.60
	GUG	187	0.59		GCG	136	0.44		GAG	330	0.51		GGG	289	0.70

Notes: The frequency of codons in the 57 chloroplast genes of *P. alba* were summed and used to determine the set of most frequently used codons. N----sum of the frequencies of the codons; RSCU----relative synonymous codon usage.

In most species examined, there is considerable heterogeneity of codon usage among genes (Ikemura 1985; Sharp et al. 1988), and so it is essential to look for any trends among genes using multivariate techniques. The correspondence analysis of relative synonymous codon usage (RSCU) is conducted and generates a series of orthogonal axes that reflect the trends responsible for the variation in codon usage. The first axis accounts for 10.37% and other three axes account for 8.80%, 8.07% and 6.94% of the total variation in the dataset, with each subsequent axis explaining a decreasing amount of the variation. Variation of 10.37% was not remarkably high for relative inertia explained by the first axis. A projection of each gene on the first two COA axes is presented in Fig. 1. The origin represents the average RSCU for all genes, with respect to the two axes. The distance between genes on this plot is a reflection of their dissimilarity in RSCU, with respect to the first two axes. The corresponding distribution of synonymous codons (Fig. 1B) shows the separation of C/G-ending codons and A/U- ending codons along the primary axis. This indicated that the variations in synonymous codon usage among the *P. alba* genes were based on the nucleotide content of the genes. The separation of genes on the second axis appears to be largely related to the level of expression. Many genes that have been known or expected to be expressed at high levels in plant chloroplast genomes such as *psaA, psaB*, *psbB*, *psbC*, *psbD*, *atpA*, *atpB* and *rbcL* and *Pet* genes (Klein et al. 1986; Mullet et al. 1987; Morton 1998), are located towards one extreme of axis 2, while others lie at the other extreme of the second axis.

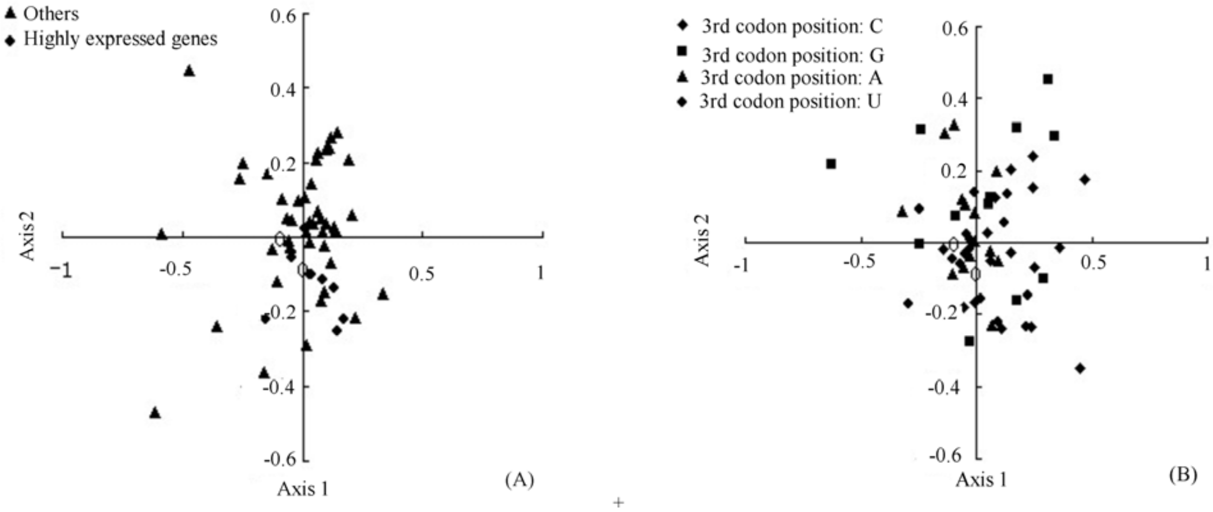

Fig. 1 Correspondence analysis of the relative synonymous codon usage in 57 genes from chloroplast genome of *Populus alba*. (A) The distribution of genes on the plane defined the first two main axes. (B) The distribution of synonymous codons along the first and the second axes of the correspondence analysis

To identify the factor that resulted in the dispersion of genes, the ordination of genes on the first four COA axes was examined for correlations with indices of codon usage and amino acid composition (e.g. ENc, GC_{3S}, GC, GRAVY and Aromaticity). A summary of these correlations is presented in Table 2. The distribution of genes on primary axis was significantly correlated with GC_{3S} (p<0.01), but was not significantly correlated with CAI (r=0.024). We can conclude that this demonstration of a strong correlation between GC_{3S} and codon usage suggests that the variation in codon usage among genes may be due to a mutational bias at the DNA level. It also was observed that the position of genes on the axis 2 and axis 3 were significantly correlated with CAI (r=-0.348, p<0.01 and r=0.602, p<0.01), meantime the ENc index was significantly correlated (r=-0.396, P<0.01) with position of genes on the third axis (axis 3). Those genes with positive coordinates on the third axis have a more biased usage of codons compared to genes with negative axis 3 coordinates. Both these observations suggest that nucleotide mutational bias plays a crucial role, selection in nature for translational efficiency only in a minor way, in shaping codon usage in the chloroplast genome of *P. alba.*

Table 2. Correlation between the codon usage and amino acid usage indices

No. of Axis	CAI	ENc	GC_{3S}	GC	Gravy	Aromaticity
Axis1	0.024	0.227	**0.349**	0.281	-0.232	-0.206
Axis2	**-0.348**	0.207	0.024	-0.146	-0.222	-0.198
Axis3	**0.602**	**-0.396**	-0.241	0.161	-0.221	0.194
Axis4	0.266	0.113	0.130	0.058	0.159	0.041

Notes: Those values that occur significantly (p<0.01) are marked in bold.

In the present study, we further investigated the relationship between nucleotide content and codon usage by ENc-plot. Wright (1990) suggested the ENc-plot (ENc plotted against GC_{3S}) as part of a general strategy to investigate patterns of synonymous codon usage. Genes, whose codon choice is constrained only by a G+ C mutation bias, will lie on or just below the curve of the predicted values (Wright 1990). The ENc-plot of genes from *P. alba* is presented in Fig. 2. It can be seen that majority of the points tracks the reference line quite closely (Fig. 2). This indicates that the observed codon bias is most easily explained as a product of G+C mutation bias. But several points with low EN_C values lay below the expected curve, including *psbA* gene which has been known that selection is acting on the codon use of this gene to adapt codons to tRNA availability in plant chloroplast genomes (Morton 1993; Pfitzinger, et al. 1987). This is probably due to the fact that codon usage is still dominated by a mutational bias in *P. alba* chloroplast genes and that selection appears to be limited to a subset of genes and to only subtly affect codon usage.

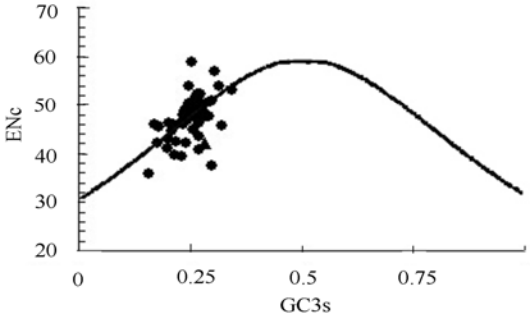

Fig. 2 Effective number of codons (EN_C) used in each gene plotted against GC content at synonymously variable third positions of codons (GC_{3S}). The continuous curve plots the relationship between ENc and GC_{3S} in the absence of selection. The *psbA* gene is indicated by a triangle

Discussion

DNA feature (base composition) is the most frequently reported and is probably one of the most pervasive influences on codon usage. Base composition is a balance between mutational pressure towards or away from G+C nucleotide pairs (Sueoka 1962). The overall genomic G+C content of *P. alba* chloroplast genome is estimated to be 36%; the mean frequency of G+C at synonymously variable third codon position (GC_{3S}) is 25.5%, somewhat lower than the estimated values for the genome. Codon usage by *P. alba* chloroplast genes is biased toward a high representation of NNU and NNT codons, similar to what have been found in other plant chloroplast genome (Wolfe et al. 1988; Morton 1993). The origin of such compositional constraints (GC/AT pressures) is still a matter of debate. Either these compositional constraints are the results of mutational biases (Sueoka 1988; Wolfe et al. 1989), or natural selection plays the major role leading to preferential fixation of non-random dinucleotide and base frequencies (Bernardi 1993; Bernardi et al. 1986; Nussinov 1984). Wolfe et al. (1992) noted that chloroplast gene codon usage appears to reflect a mutational bias rather than selection. This is also suggested by the correlation between genome AT content and overall codon bias (Morton 1993). In the plants, *Pinus thunbergii* and the flowering plants appear to have very weak natural selection on codon usage of chloroplast genome such that only a few genes, in particular *rbcL* and *psbA*, show any evidence for selection, but there is evidence that an intermediate level of selection exists in the liverwort *Marchantia polymorpha* (Morton 1998).

In the present study, correspondence analysis and ENc-plot are used to investigate patterns of synonymous codon usage of *P. alba* chloroplast genome. The demonstration of a strong correlation between the frequency of use of G+C in synonymously variable third positions of sense codon (GC_{3S}) and codon usage suggested that the variation in codon usage among genes might be due to a mutational bias at the DNA level rather than natural selection acting at the level of mRNA translation. Although codon usage of chloroplast genes appears to be a result of a mutational bias toward a high AT content, the data presented here also suggest that there is some selection acting on codon usage of a few *P. alba* chloroplast genes. The evolution of codon bias over all of the chloroplast lineages is a complex matter. Several factors are likely to be involved in determining the selective constraints on codon bias (Morton 1998), and recent work has indicated that it is a dynamic process (Morton et al. 1997). The variation in selective constraints among the different lineages also makes it likely that substitution dynamics are substantially different in different lineages which might be related to the debate concerning how composition bias influences the phylogenetic reconstruction of chloroplast origins (Lockhart et al. 1992).

References

Akashi H. 2001. Gene expression and molecular evolution. *Curr Opin Genet Dev*, **11**: 660–666.

Bernardi G. 1986. Compositional constraints and genome evolution. *J Mol Evol*, **24**: 1–11.

Bernardi G. 1993. The vertebrate genome - isochores and evolution. *Mo Biol Evol*, **10**: 186–204.

Bonitz SG, Berlani R., Coruzzi G, Li M, Macino G. 1980. Codon recognition rules in yeast mitochondria. *Proc Natl Acad Sci U.S.A.*, **77**(6): 3167–3170.

Duret L. 2002. Evolution of synonymous codon usage in metazoans. *Curr Opin Genet. Dev.*, **12**: 640–649.

Grantham R, Gautier C, Gouy M, Mercier R, Pave A. 1980. Codon catalogue usage and the genome hypothesis. *Nucleic Acids Res*, **8**: r49–r62.

Hallick RB, Bairoch A. 1994. Proposals for the naming of chloroplast genes. III. Nomenclature for open reading frames encoded in chloroplast genomes. *Plant Mol Biol Reptr*, **12**: S29–30.

Ikemura T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol*, **2:** 13–34.

Kawabe A, Miyashita NT. 2003. Patterns of codon usage bias in three dicot and four monocot plant species. *Genes Genet. Syst*, **8**: 343–352.

Klein RR, Mullet JE. 1986. Regulation of chloroplast-encoded chlorophyll-binding protein translation during higher plant chloroplast biogenesis. *J. Biol. Chem*., **26**1: 11138–11145.

Lockhart PJ, Penny D, Hendy MD, Howe CJ, Beanland TJ, Larkum AWD. 1992. Controversy on chloroplast origins. *FEBS Lett*., **301**: 127–131.

Lu H, Zha, WM, Zheng Y, et al. 2005. Analysis of synonymous codon usage bias in *Chlamydia*. *Acta Biochimica et Biophysica Sinic*, **37**(1): 1–10.

Morton BR. 1993. Chloroplast DNA codon use: evidence for selection at the psbA locus based on tRNA availability. *J Mol Evol*, **37**:273–280.

Morton BR. 1998. Selection on the codon bias of chloroplast and cyanelle

Genes in Different Plant and Algal Lineages. *J Mol Evol*, **46**: 449−459.

Morton BR. 1999. Strand asymmetry and codon usage bias in the chloroplast genome of Euglena gracilis. *Proc Natl Acad. Sci USA*, **96**(9): 5123–5128.

Morton BR. 2003. The role of context-dependent mutations in generating compositional and codon usage bias in grass chloroplast DNA. *J Mol Evol*, **56**(5): 616−29.

Morton BR, Levin JA. 1997. The atypical codon usage of the psbA gene may be the remnant of an ancestral bias. *Proc Natl Acad. Sci. USA*, **94**: 11434−11438.

Morton BR, So BG. 2000. Codon usage in plastid genes is correlated with context, position within the gene, and amino acid content. *J Mol Evol*, **50**: 184−193.

Mullet JE, Klein RR. 1987. Transcription and RNA stability are important determinants of higher plant chloroplast RNA levels. *EMBO J*, **6**: 1571−1579.

Nussinov R. 1984. Strong doublet preferences in nucleotide sequences and DNA geometry. *J Mol Evol*, **20**: 111−119.

Perriere G, Thioulouse J. 2002. Use and misuse of correspondence analysis in codon usgae studies. *Nucleic Acids Res*, **30**: 4548−4555.

Pfitzinger H, Guillemaut P, Weil JH, Pillay DTN. 1987. Adjustment of the tRNA population to the codon usage of chloroplasts. *Nucleic Acids Res*, **15**: 137.

Sharp PM, Cowe E, Higgins DG, Shields DC, Wolfe KH, Wright F. 1988. Codon usage in Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster and Homo sapiens; a review of the considerable within-species diversity. *Nucleic Acids Res*, **16**: 8207−8711.

Sharp PM, Devine KM. 1989. Codon usage and gene-expression level in Dictyostelium discoideum - highly expressed genes do prefer optimal codons. *Nucleic Acids Res.*, **17**: 5029−5039.

Sharp, PM, Li, WH. 1987. The codon adaptation index─A measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res*, **15**(3): 1281−1295.

Sharp PM, Tuohy TMF, Mosurski KR. 1986. Codon usage in yeast cluster-analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res*, **14**: 5125−5143.

Shields DC, Sharp PM. 1987. Synonymous codon usage in Bacillus subtilis reflects both translational selection and mutational biases. *Nucleic Acids Res*, **15**: 8023−8040.

Sueoka N. 1962. On the genetic basis of variation and heterogeneity of DNA base composition. Proc. *Natl Acad Sci USA*, **48**: 582−592.

Sueoka N. 1988. Directional mutational pressure and neutral pressure. *Proc. Natl Acad Sci USA*, **85**: 2653−2657.

Wang HC, Hickey DA. 2007. Rapid divergence of codon usage patterns within the rice genome. *BMC Evol Bio*, **7** (Supp 1): S6.

Wolfe KH, Morden CW, EMS SC, Palmer JD.1992. Rapid evolution of the plastid translational apparatus in a nonphotosynthetic plant: loss or accelerated sequence evolution of tRNA and ribosomal protein genes. *J Mol Evol*, **35**: 304−317.

Wolfe KH, Sharp PM. 1988. Identification of functional open reading frames in chloroplast genomes. *Gene*, **66**: 215−222.

Wolfe KH, Sharp PM, Li WH. 1989. Mutation-rates differ among regions of the mammalian genome. *Nature*, **337**: 283−285.

Wright F. 1990. The ‘effective number of codons’ used in a gene. *Gene*, **87**(1): 23−29.